

DEPARTMENT OF ECONOMICS AND FINANCE
COLLEGE OF BUSINESS AND ECONOMICS
UNIVERSITY OF CANTERBURY
CHRISTCHURCH, NEW ZEALAND

**TO USE CONSTRUCTED-RESPONSE QUESTIONS, OR NOT
TO USE CONSTRUCTED-RESPONSE QUESTIONS? THAT IS THE
QUESTION**

Stephen Hickson, W. Robert Reed, and Nicholas Sander

WORKING PAPER

No. 69/2010

**Department of Economics and Finance
College of Business and Economics
University of Canterbury
Private Bag 4800, Christchurch
New Zealand**

TO USE CONSTRUCTED-RESPONSE QUESTIONS, OR NOT TO USE CONSTRUCTED-RESPONSE QUESTIONS? THAT IS THE QUESTION

by

Stephen Hickson, W. Robert Reed, and Nicholas Sander*

Abstract

Advocates of Constructed Response (CR) questions argue that CR questions provide a different assessment of student knowledge than is available from Multiple Choice (MC) questions. If that is the case, and if the benefit in terms of improved assessment is substantial, then it follows that grade outcomes using CR questions should be different from those using MC questions. We investigate this using a large dataset composed of individual assessment results from thousands of students in introductory economics classes at a large public university. Empirical analysis of our large sample of students indicates that a switch to an all-MC format would result in grade changes that are in the “small” to moderate range when compared to grade changes that occur between assessments. This evidence suggests that CR questions could be abandoned at relatively little cost in grading accuracy. However, there are other arguments in favour of keeping CR questions. In particular, it has been suggested that students perceive a mix of CR and MC as “fairer” than an assessment composed exclusively of one or the other question type. Further, some instructors believe that CR questions encourage students to study harder. We provide survey evidence that supports both arguments.

JEL Categories: A22

Keywords: Principles of Economics Assessment, Multiple Choice, Constructed Response, Free Response, Essay.

October 19, 2010

*Hickson, Reed, and Sander are, respectively, Teaching Fellow, Professor, and Honours student at the University of Canterbury. Sander was supported by a Summer Scholarship 2009-2010 jointly funded by the University of Canterbury and the Tertiary Education Commission of New Zealand (TEC). Reed is the contact author and his contact details are: Department of Economics and Finance, University of Canterbury, Private Bag 4800, Christchurch 8042, New Zealand; Email: bobreednz@yahoo.com; Phone: +64 3 364 2846.

“A main message of this article is that the decision to abandon the constructed-response portion of tests in favour of an all-multiple-choice format should rest on the analysis and considered judgment of the costs and benefits of such a decision.”

- Kennedy and Walstad (1997)

I. INTRODUCTION

Most principles classes in economics employ a variety of question formats in assessing students. These can be categorized into two broad categories. Multiple choice (MC) questions present students a set of answers and ask them to select the correct one(s). Constructed response (CR) questions require students to provide their own answers. These can range from fill-in-the-blank questions; to definitional or short-answer questions; to questions requiring mathematical solutions; to long essay questions. CR questions can be costly to evaluate, and are subject to greater subjectivity in marking. However, many instructors believe that (i) CR questions measure a different type of knowledge/learning than MC questions, and (ii) that this benefit is sufficiently large to outweigh their higher costs.

The research literature is divided with respect to the first of these points. Many studies conclude that MC and CR questions measure the same thing (Bennett, Rock, and Wang, 1991; Lukhele, Thissen, and Wainer, 1994; Thissen, Wainer, and Wang, 1994; Wainer and Thissen, 1993; and Walstad and Becker, 1994). Other studies find evidence that MC and CR measure different things (Krieg and Uyar, 2001; Lumsden and Scott, 1987; Becker and Johnston, 1999; and Hickson and Reed, 2009).

Even if CR questions measure a different type of knowledge/learning, there remains the question of whether this additional information makes much of a difference in terms of assessment outcomes or grades. If CR questions produce the same grades as MC questions – or near to the same grades – then there may be little benefit to using CR questions.

Accordingly, the starting point for this study is an investigation of the extent to which grades based solely on MC questions differ from grades based solely on CR questions. While other studies have focused on test scores, frequently AP test scores, our study is the first to focus attention on grades. Further, the data for our study is collected directly from university introductory economics classes. This should make our findings of particular interest to instructors of large, university classes who are considering switching to all-MC assessments.

We first develop a theoretical framework for analyzing MC-CR grade differences, decomposing observed grade differences into (i) systematic and (ii) sampling error components. We show that, while it is generally impossible to identify the separate influences of these two components, MC-CR grade differences can still provide useful information. Specifically, they can be used to benchmark the size of grade differences that might be expected from switching to all-MC assessments.

Our empirical analysis employs data on thousands of individual students who took principles of economics classes at a large public university. We develop an empirical procedure for estimating the differences in grades that students would receive if their grade was based solely on either the MC- or the CR-question portion of their assessments. As we detail below, the instructors of these classes made a conscientious effort to write CR questions so that they would assess higher-level learning missed by MC questions (Bloom, 1956). Thus, these data provide an excellent opportunity to evaluate whether CR questions produce different grade outcomes than MC questions.

Our analysis finds that the grade differences between CR and MC questions on a given assessment are similar in distribution to (i) differences in CR grades across different assessments, and (ii) differences in MC grades across different assessments. In other words, switching to all-MC assessments, holding other things constant, would not introduce a greater change in students' grades than students currently experience across assessments. The effect

would be even smaller if assessments currently use a mix of CR and MC, so that the switch was from a composite test to an all-MC assessment, rather than from an all-CR format to an all-MC format. If the resulting grade differences are not very large, this suggests that the benefits of using CR questions may also not be very large.

However, there may be benefits to using CR questions that are not captured by the preceding empirical analysis. These have to do with (i) fairness and (ii) incentive effects on students' study behaviors. To pursue this further, we survey a smaller number of students regarding their preferences for MC and CR questions. While many students believe they would personally benefit from assessments that were exclusively composed of either all-MC or all-CR questions, the great majority of students believe that the fairest test is one that uses a mix of these questions. Finally, when asked which type of assessment would induce them to study harder, most students say they would study the same no matter which question format was used. However, a substantial minority of students say they would study harder for a test consisting of CR questions, with almost no students stating they would study harder if the test were composed of all-MC questions.

The paper proceeds as follows. Section II describes our data. Section III provides the theoretical framework for understanding the empirical analysis. Section IV presents our results. Section V describes the type of MC and CR questions underlying our data. Section VI concludes.

II. DATA

Data. Our data consist of 7754 assessment records for students who took introductory microeconomic and/or introductory macroeconomics at the University of Canterbury (UC) from 2002-2007. For each student in each class we have a record of both their (i) Term Test and (ii) Final Exam results, including their performance on the MC and CR components of

each assessment.¹ Term Tests for both micro- and macroeconomics classes consisted of a MC component consisting of 25 questions worth one point each, and a CR component consisting of 2 questions worth a total of 50 points. Final Exams consisted of 30 MC questions worth one point each, and 3 CR questions worth a total of 70 points. The same two instructors taught both courses over this time period.

We converted scores for each of the four components – MC-Term Test, MC-Final Exam, CR-Term Test, CR-Final Exam – to a 100-point scale. We then converted these numerical scores to grades using the UC grading scheme: A+ is assigned to scores 85 and over; A to scores between 80 and 85, A- to scores between 75 and 80, ... , C- to scores between 45 and 50, D to scores between 40 and 45, and E to all scores less than 40. The respective grade distributions for each of the four assessment components are reported in TABLE 1, as are some other assessment characteristics.

The bottom two rows in the table report the mean and standard deviation of raw assessment scores (adjusted to a 100-point scale) for each of the assessment components. The first four columns report the distribution of grades that would be given if each of the components constituted a separate assessment. The last column reports the historical distribution of final grades given to the students in our sample.

The first thing we note is that students in our sample typically scored higher marks on the MC components of the assessments in our sample. The mean scores of the MC components were 66.5 and 70.3 on the Term Tests and Final Exams, respectively. The corresponding means for the CR components were 49.5 and 53.4. The higher mean scores for the MC components translate to higher grades. If the MC components were stand-alone assessments, students would receive more As and Bs, and fewer Cs, Ds, and Es. In other

¹ We excluded students who did not complete both a term test and a final exam. Likewise, we excluded all students who were awarded an “aegrotat” as this flagged assessment results that were deemed unrepresentative due to circumstances which substantially impaired a student’s performance.

words, if the grading schedules were kept the same and everything else remained constant, students would see their grades increase substantially if assessments switched to all-MC questions. This highlights a major problem with measuring the impact of a greater reliance on MC questions.

If MC questions tend to produce, say, higher scores than CR questions, we suspect that instructors would either write harder MC questions, or scale the marking scheme to make it more difficult to achieve higher grades. We incorporate this response in our analysis by assuming that any change in the question makeup of assessments will maintain the overall grade distribution. This requires that we adopt a pre-determined grade distribution to standardize the grade outcomes from MC and CR questions. We operationalize this by using the distribution of actual final grades for all students in our sample. This distribution is reported in the last column of TABLE 1.

We use this historical distribution to “norm” individual assessment components by imposing this grade distribution on each of the respective components. For example, for a given class with say, 400 students, we have the following four assessment observations for each student: a MC(Term Test), a MC(Final Exam), a CR(Term Test), and a CR(Final Exam). For each assessment component we assign A+s to the top 3.4 percent of students, A’s to the next 4.7 percent of students, and so on. In this way, the same grade distribution is imposed on each assessment component for each class in our sample. Thus, when we take the difference between the MC and CR components of an assessment, we hold constant the overall grade distribution.

TABLE 2 maps students’ MC grades conditional on their CR grades on the same assessment. The top panel reports grade distributions for Term Tests. The bottom panel does the same for Final Exams. The table is interpreted thusly: 28.5 percent of students who received an A+ on the CR component of Term Tests also received an A+ on the MC

component (once we rescaled CR and MC scores to have the same grade distributions). 17.2 percent received an A on the MC component. And so on.

Along any given row in the table, the sum of percentages excluding the diagonal element represents the percent of students who would experience a grade change if their assessment grade was based entirely on their MC component rather than their CR component. Thus, 72.5 percent of students ($=100 - 28.5$) who received an A+ on the CR component of the Term Test would receive a different grade (in this case, a lower grade) if their grade was based on the MC component. The numbers in the table can also be used to calculate the percentage of students who experienced a change of one full grade (e.g., B+ to C+, B- to A-) or more.

TABLE 2 is insightful for assessing how the rank order of students changes when grades are based on MC questions rather than CR questions. Small changes in rank order are likely to show up as no change in grade. In contrast, rank orders must be substantially altered in order for a change in question format to produce large grade changes. The next section provides a theoretical framework for analyzing observed grade differences.

III. THEORY

Grade differences between MC and CR components on the same assessment. Let an individual observation of student i 's MC component on a given assessment j (j = Term Test, Final Exam) be given by

$$(1) \quad y_{ij,MC} = \mu_{ij,MC} + \varepsilon_{ij,MC},$$

where $y_{ij,MC}$ is the observed grade measured in points (cf. TABLE 1), $\mu_{ij,MC}$ is the mean grade the i th student would achieve on the j th MC assessment, and $\varepsilon_{ij,MC}$ is the “sampling error” associated with the fact that the same student would score different marks on different offerings of a given MC assessment. “Sampling error” may be due to student-specific

factors, such as how hard the student studied for a particular assessment, how mentally alert and focused the student was on a given day, etc. Or it may be due to instructor-specific factors such as the general difficulty of the questions posed on the given assessment, or the specific areas that the questions tested. “Sampling error” also includes the impact of guessing.

Note that the systematic component, $\mu_{ij,MC}$, may incorporate factors that are not directly related to an understanding of course material. For example, some students may generally experience less test anxiety than others. This may result in systematically higher grades, even though the student may not have a superior understanding of the material.

Similarly, let

$$(2) \quad y_{ij,CR} = \mu_{ij,CR} + \varepsilon_{ij,CR}$$

represent individual i 's CR grade on the j th assessment, where $\mu_{ij,CR}$ and $\varepsilon_{ij,CR}$ are defined as above, except that $\varepsilon_{ij,CR}$ also includes “marking error” associated with subjective evaluations by potentially different markers evaluating the student's CR answers.

It follows that the observed difference between a student's MC and CR grades on a given assessment equals

$$(3) \quad y_{ij,MC} - y_{ij,CR} = (\mu_{ij,MC} - \mu_{ij,CR}) + (\varepsilon_{ij,MC} - \varepsilon_{ij,CR}).$$

This difference is composed of two components: (i) a “systematic” difference, $(\mu_{ij,MC} - \mu_{ij,CR})$, which represents the expected value of the difference in scores from a student repeatedly taking CR and MC assessments covering the same material; and (ii) “sampling error,” $(\varepsilon_{ij,MC} - \varepsilon_{ij,CR})$, associated with a variety of student- and instructor-specific factors.

If CR questions are able to assess higher-level learning in a way that MC questions are not, this should be reflected in the systematic component, $(\mu_{ij,MC} - \mu_{ij,CR})$. If the systematic component shows little variation, then it would call into question the benefit of using CR questions.² Accordingly, we would like to be able to identify the systematic component in observed MC-CR grade differences.

FIGURE 1 illustrates the problem. The first two panels show the distributions of MC and CR grades for a given assessment, measured in points.³ Both distributions are normed to have the same overall grade distribution. For each student and each assessment, we match their grades from the MC and CR components of that assessment and take the difference. The two lower panels of FIGURE 1 plot the distribution of MC-CR grade differences for different cases that we examine below. Note that the values on the horizontal scale run from -10 to +10. “-10” represents the grade difference associated with a student receiving an A+ and an E on the CR and MC components of an assessment, respectively. “+10” represents the grade difference from receiving an E on the CR component and an A+ on the MC component of that assessment. Consider the following cases:

Case One. Case One represents no sampling error and no systematic differences, $(\varepsilon_{ij,MC} - \varepsilon_{ij,CR}) = (\mu_{ij,MC} - \mu_{ij,CR}) = 0$, for all i, j . In this case, all students receive the same grade on the MC and CR components, so that the distribution of $y_{ij,MC} - y_{ij,CR}$ is represented by a mass point at 0.⁴ In other words, the rank order of students is the same for both MC and CR.

² This systematic component may incorporate the influence of factors not directly related to understanding. For example, non-native English speakers may have a more difficult time expressing their thoughts on CR questions. Markers may not be able to distinguish (nor care to distinguish) that an unsuccessful answer is due to poor English facility rather than poor comprehension of course material. Of course, some may argue that it is desirable for students to be able to express themselves in English, so that it is desirable to capture this effect in the systematic component.

³ For the relationship between grades and points, see TABLE 1.

⁴ In terms of TABLE 2, all the diagonal terms would equal 100 and the off-diagonal terms would be zero.

Case Two. Case Two occurs when there is no sampling error but systematic differences exist between students' performances on the MC and CR components of assessments: $(\varepsilon_{ij,MC} - \varepsilon_{ij,CR}) = 0$, $(\mu_{ij,MC} - \mu_{ij,CR}) \neq 0$ for all i,j . Case Two will produce distributions of grade changes like those in Panel (D) of FIGURE 1. In this case, the distribution of $y_{ij,MC} - y_{ij,CR}$ contains useful information because it measures the systematic effects of switching from an all-CR to an all-MC question format.

The near symmetry of the grade difference distribution in Panel (D) is no accident. Given the fact that the MC and CR grade distributions are normed to be the same, the total number of grade point increases must equal the total number of grade point decreases. For example, if one student experiences a grade increase from an B- to a B+, either (i) another student must experience a grade decrease from a B+ to a B-, or (ii) two students must experience a grade change where one decreases from B+ to B, and another decreases from B to B-, or (iii) a similar chain of grade changes must take place to compensate the original change.

Case Three. Case Three represents the scenario where there are no systematic differences between students' performances on the MC and CR components of assessments, but sampling errors cause the observed grades to be different, $(\mu_{ij,MC} - \mu_{ij,CR}) = 0$, $(\varepsilon_{ij,MC} - \varepsilon_{ij,CR}) \neq 0$ for all i,j . This will also produce distributions of grade differences like those reported in Panel (D) of FIGURE 1. However, in this case, the difference distribution tells us nothing about the systematic effects of switching from CR to MC assessments. Unfortunately, there is no way to distinguish Case Two from Case Three as they are observationally identical.⁵

⁵ Note that the norming of the MC and CR distributions throws away information about the relative sizes of the sampling errors associated with the original test scores. This is a direct consequence of assuming that

Case Four. In Case Four, both systematic differences and sampling errors are jointly present, $(\varepsilon_{ij,MC} - \varepsilon_{ij,CR}) \neq 0$, $(\mu_{ij,MC} - \mu_{ij,CR}) \neq 0$ for all i, j . This will produce distributions like the two previous cases. While this is the case that most certainly represents reality, it is observationally equivalent to the two previous cases. Without imposing an assumption on their relative sizes, it is impossible to identify the systematic component of the MC-CR grade differences.⁶

In fact, the problem of identifying systematic differences from observed differences is even more vexing than the previous discussion acknowledges. Note that the size of the systematic component, $(\mu_{ij,MC} - \mu_{ij,CR})$, in the grade difference distribution of Panel (D) is not independent of the behavior of the error terms. As the variance of the sampling errors ε increases, so must the variance of the corresponding grade point distributions y . These need to get re-normed in order to maintain the overall grade point distribution. This effectively reduces the variance of the systematic component μ , and thus, similarly, the variance of $(\mu_{ij,MC} - \mu_{ij,CR})$.

The preceding discussion illuminates the problems of identifying the effects of switching to an all-MC assessment format based on observed differences in MC and CR scores. The remainder of this section discusses how observed MC and CR differences can be analyzed to achieve a more modest goal that may still provide useful information regarding the benefits of CR questions.

instructors would use similar/identical grade distributions irrespective of whether the assessments were MC or CR.

⁶ Kennedy and Walstad (1997) get around this problem because they (i) use nominal scores, rather than a normed distribution; and (ii) assume the size of the variance of the error terms in their sample of AP scores based on external analyses, i.e., they do not estimate it from their sample.

Grade differences between similar components across different assessments. Let us now consider differences in MC grades between the Term Test (j) and Final Exam (k) for a given student i . Using the same notation as above, the observed difference is given by

$$(4) \quad y_{ij,MC} - y_{ik,MC} = (\mu_{ij,MC} - \mu_{ik,MC}) + (\varepsilon_{ij,MC} - \varepsilon_{ik,MC}).$$

A similar relationship holds for the differences in CR grades across different assessments j and k .

$$(5) \quad y_{ij,CR} - y_{ik,CR} = (\mu_{ij,CR} - \mu_{ik,CR}) + (\varepsilon_{ij,CR} - \varepsilon_{ik,CR}).$$

Once again the observed difference is composed of two components: (i) a “systematic” difference, and (ii) “sampling error.”

Note, however, that the two components differ in important ways from the two components in Equation (3). While MC and CR questions on the same test may assess different parts of course material, the overlap in course material is likely to be greater than the overlap on the Term Test and Final Exam. Further, the sampling error component now includes differences across different points in time that are not related to understanding. For example, a student may feel fine for the Term Test but feel sick for the Final Exam. Nevertheless, all of the problems associated with inferring the sizes of these two components from observed MC-CR differences hold *a fortiori* when analyzing MC-MC and CR-CR differences across assessments.

The value of cross-assessment differences in MC and CR grades. Given the above it is clear that one cannot use MC-MC or CR-CR differences across assessments to identify the systematic component of the distribution of MC-CR grade differences on the same assessment. Nevertheless, differences across assessments can still provide useful information.

The greater the dispersion in MC-CR grade differences, the greater the potential for a switch to all-MC questions to substantially alter grade outcomes. But how can one determine

whether a distribution of MC-CR grade differences is “large” or “small?” Students and instructors are accustomed to the fact that students’ class ranks vary across assessments. We argue that this provides a natural benchmark against which to judge the size of MC-CR grade differences.

Suppose the change in the rank order of students that occurs from moving from an all-CR to an all-MC format were substantially larger than that which exists between the Term Test and Final Exam. While the larger differences could be due to sampling error, there is the possibility that these differences are picking up systematic differences between what MC and CR are measuring. If instructors were inclined to believe that CR questions measure a different level of understanding than MC questions, then this “large” difference in grade outcomes could indicate that there was a significant cost to switching to an all-MC assessment.

On the other hand, suppose the change in the rank order of students that occurs from moving to an all-MC format was substantially smaller than between-assessment grade differences. That would suggest that CR questions produced grade outcomes that were very similar to those from MC questions. In this case, instructors could switch to the lower-cost MC questions with little cost to assessment accuracy.

IV. RESULTS

The first two rows of TABLE 3 summarize grade differences between the MC and CR components on Term Tests and Final Exams, respectively. We categorize the differences in terms of whether there is (i) a one-letter grade negative difference or more between the CR and MC components, (ii) a difference of less than one-letter grade, and (iii) a one-letter grade positive difference or more. An example of a negative, one-letter grade difference would be a student who received a grade of A- on the CR component while achieving a B- on the MC component. That is, their MC grade was a letter grade lower than their CR grade.

For the Term Test observations in our sample, roughly one-third of students had a MC-CR grade difference of a letter grade or more ($32.9\% = 16.2\% + 16.7\%$). For the Final Exam observations, the corresponding number is a little less than a fourth ($22.7\% = 11.4\% + 11.3\%$). As discussed above, it is unclear how to map this empirical result to the decision to move to all-MC assessments. One way to measure the size of these grade differences is to compare them with between-assessment grade differences.

The third and fourth rows of TABLE 3 summarize the differences between MC grades on the Term Test and Final Exam, and CR grades on the Term Test and Final Exam, respectively. Interestingly, there is a greater difference in MC grades across assessments than in CR grades. Roughly a third of students saw their MC grade change by a letter grade or more between the Term Test and Final Exam. Approximately a fourth of students saw their CR grades change by a letter grade or more.

On the face of it, the MC-CR differences are approximately equal in size to grade differences that occur between assessments. However, it should be remembered that the MC-CR grade differences represent the change in grade outcomes that would arise from switching from an all-CR assessment to an all-MC assessment. The classes analyzed in this study relied on a mix of CR and MC questions. The effect of switching from this composite format to an all-MC would be proportionately less. If we use between-assessment grade differences as our benchmark, these results suggest that the switch to an all-MC format would result in grade changes that were in the “small” to moderate range.

Does this mean that CR questions could be abandoned at relatively little cost in grading accuracy? Two other arguments that are sometimes given in favour of maintaining a composite MC/CR question format are that (i) it is “fairer” than relying solely on MC questions, and (ii) using CR questions encourages students to study harder. To address these issues, the authors surveyed 131 students in a Principles of Macroeconomics class about their

attitudes towards MC and CR questions. The questions and responses are reported in TABLE 4.

The first question asks students whether they think they would do better with an all-MC assessment, an all-CR assessment, or a mixture of the two. Approximately 15 percent of students said they would do better with an all-MC assessment. This was matched by an almost identical percentage of students who thought they would do best with an all-CR assessment. While the set of choices for this question do not map directly onto the results of TABLE 3, there is an interesting similarity. From the first two rows of TABLE 3, we see that 16.2% and 11.4% of students would have experienced at least a letter grade advantage if their grade was based on CR questions rather than MC questions. On the other side, 16.7% and 11.3% of students would have been at least a letter grade better off with MC questions. These numbers generally confirm the self-assessment results of TABLE 4, with approximately equal numbers of students seeing themselves as being better off with one type of question as opposed to the other.

The second question in TABLE 4 addresses the fairness issue. Despite the fact that substantial minorities see themselves as being personally advantaged by one question-type or the other, almost 80 percent of students state that the fairest assessment would be one that relied on both types of questions. Interestingly, the 39 students who expressed a preference for all-MC or all-CR questions in Question 1 have very similar opinions as to what type of test format is fairest. Despite their personal, self-perceived advantage from one or the other question type, they also think the fairest test would be one that had a mix of both questions.

Unknown is the extent to which students' notions of fairness are conditioned on an expectation that grade outcomes would be very different if assessments were composed exclusively of MC or CR questions. It is possible that students would evaluate the fairness

issue differently if they thought that a change in question format had only a small effect on grade outcomes.

The third question addresses the incentive effects of the two question types. Very few students say they would study harder if assessments were composed entirely of MC questions. A little over a third say they would study harder if the assessment has CR questions, with the remainder saying that question format would not affect how hard they studied.

V. A CLOSER LOOK AT THE MC AND CR QUESTIONS UNDERLYING THIS STUDY

It is difficult to assess the contribution of our study without an appreciation of the kinds of MC and CR questions that were used in the classes we studied. The instructors of these classes made a conscientious effort to write CR questions so that they would assess higher-level learning missed by MC questions (Bloom, 1956). This section attempts to illuminate the generally different natures of MC and CR questions as they were employed in the assessments investigated in this study. Before doing so, it is helpful to briefly review the literature on the ability of MC and CR questions to measure higher-order learning outcomes.

Bloom (1956) defines the following six levels of learning (our expanded explanations are in parentheses);

1. Knowledge (knowing facts);
2. Comprehension (understanding the importance of known knowledge);
3. Application (putting knowledge and understanding to use);
4. Analysis (using knowledge to breaking down a problem into component parts);
5. Synthesis (combining different parts to form new knowledge and ideas); and
6. Evaluation (determining the worth or usefulness of knowledge, application, analysis or synthesis).

Textbook, MC test banks tend to consist of questions that disproportionately sample from the first two levels of learning. Buckles and Siegfried (2006) conclude that MC questions can be effectively used to assess up through the first four levels of Bloom's

taxonomy. In contrast, they argue that while it is possible to use MC questions to assess Synthesis and Evaluation, these are more reliably measured through CR questions. According to Buckles and Siegfried (2006), the key ingredient for assessing these higher-level learning outcomes is the requirement that students work through a chain of reasoning using a number of logical steps. It is difficult to write a sequence of MC questions that get at this learning dimension, especially when the chain of reasoning can involve a complicated decision tree.

These conclusions find support elsewhere in the literature. As part of a wider study, Iz and Fok (2007) attempt to classify the set of 25 MC questions used in the test for the Higher Diploma of Surveying. They classify 21 of the 25 as levels 1 to 4. The remaining four questions were simply lumped together as “they were few in numbers... and difficult to discriminate”. Zheng et al (2008) assert that it is “...much more difficult to write multiple-choice questions at the Application and Analysis levels of Bloom’s taxonomy than at the Knowledge or Comprehension levels.” It is even more difficult to write Synthesis and Evaluation MC questions. Thus it is no surprise that standard textbook question banks are dominated by recognition-, recall-, and understanding-type questions.

Walstad (2006) concurs with Buckles and Siegfried to a large extent, but notes that many CR questions are not well-designed to assess higher-level learning. Unless they are carefully constructed, CR questions may only be testing recall and recognition. A key issue is whether the student could have memorized the answer in advance.

We next describe the nature of the MC and CR questions used in the assessments included in our data set.⁷ MC and CR were deliberately constructed to assess different levels

⁷ The questions are taken from the term-test and final exam for Introduction to Macroeconomics (ECON 105), Semester One, 2006.

of knowledge. The first example is a MC question that was designed to test for Knowledge (Level 1 of Bloom's taxonomy).

Which of the following is NOT an impact of inflation?

- 1. Wealth is transferred from savers to borrowers.*
- 2. Important price signals become more difficult to read.*
- 3. The currency loses value.*
- 4. The value of money assets rises.*

The next example is another MC question, but this one was designed to test for Application and Analysis (Levels 3 and 4).

A recession in the rest of the world is likely to cause _____ GDP growth and _____ inflation in New Zealand.

- 1. higher; higher.*
- 2. higher; lower.*
- 3. lower; higher.*
- 4. lower; lower.*

Assessing higher levels of knowledge becomes much more difficult with MC questions. This is where CR questions provide an opportunity to assess levels of knowledge that cannot, or at least are not, being measured by MC questions.

The following example is taken from the same course as the questions above. It illustrates how a CR question can be written such that higher levels of learning are progressively tested as the student works their way through the question.

In 1989, the Government passed the Reserve Bank Act. How would you characterise the NZ economy since that time in terms of growth, inflation and unemployment?

This question tests Knowledge and Comprehension (Levels 1 and 2). It could be easily rewritten in a MC format. Marks were awarded for stating how economic growth, inflation and unemployment had performed over this period in general terms (Knowledge). Marks were also awarded for answers that commented on the importance of these facts (e.g. recent slowing of growth at that time).

A follow-up CR question is:

The Reserve Bank Monetary Policy news release above [not shown here] was issued on 9 March 2006. In this release the Bank identifies a number of factors that are influencing both inflation and growth. Use an AD/AS model to explain how the Reserve Bank currently sees the following factors influencing inflation and growth (remembering that the AD/AS model is a static model so you will need to interpret the results).

- (i) the slowing (or cooling) of the housing market.*
- (ii) labour costs.*
- (iii) business confidence.*

This question tests Application, Analysis and some Synthesis (Levels 3, 4, and 5). Students are required to break down the economic factors identified in the Reserve Bank news release and to use the AD/AS model to analyse the question. The student needs to have a good working knowledge of the AD/AS model because the question does not explicitly identify how AD/AS are affected by the respective factors. Further, the student must bring these factors together to determine their overall impact on growth and inflation. The latter involves extending results from the static model (price and GDP level) to a dynamic world (inflation and growth).

The next CR question succeeds the previous one and moves to Synthesis and Evaluation (Levels 5 and 6):

If the three influences analysed above were the only factors impacting the NZ economy, what conclusions would you make about the outlook for inflation and growth?

Students must combine all three answers into one overall judgement. From the answers to the previous question there is no ambiguity about the impact on economic growth but the impact on inflation of these three influences is ambiguous. Students need to recognise this and answer accordingly. The question and the resources provided with the question contain little guidance for the student. Further, students must provide a consistent answer based on their previous answer.

Typically, students who have learnt some facts will achieve a good score on the first CR question. Students who have learnt the mechanics of the AD/AS model will earn at least

some of the marks for the second CR question. The most able students will earn marks for the last CR question.

These latter examples are designed to illustrate the difficulty with writing MC questions to assess the highest levels of learning. These levels of learning are best assessed when the student is asked to analyze a complex economic question that requires them to assemble a chain of logical arguments. Consider the problem of assessing such a problem with MC question(s). If a single MC question is used to assess a problem of great complexity, fairness would dictate that it be worth many more points than simple recognition, MC questions. But the all-or-nothing marking of MC questions makes this a risky measure. In contrast, if a sequence of MC questions is used to assess the different parts of the logical chain, it is difficult to not lead the student into the answer by virtue of asking the question(s). The combination of their free-response nature, along with partial-credit marking, endows the CR question format with the potential to better assess higher-level learning while maintaining fairness to students.

In conclusion, MC and CR questions are most likely to produce different outcomes when an intentional effort is made to use them to assess different levels of knowledge. Such was the case in the introductory economics classes from which our sample was drawn. If substantial grade differences were not observed here, then it is unlikely that they will be observed when such an effort is not made.

VI. CONCLUSION

Advocates of Constructed Response (CR) questions argue that CR questions provide a different assessment of student knowledge than is available from Multiple Choice (MC) questions. If that is the case, and if the benefit in terms of improved assessment is substantial, then it follows that grade outcomes using CR questions should be different from those using MC questions. We investigate this using a large dataset composed of individual

assessment results from thousands of students in introductory economics classes at a large public university.

Our theoretical framework demonstrates that one must be careful in interpreting MC-CR grade differences. These differences are composed of two parts: (i) a systematic component, which should pick up the benefits of improved grading accuracy from using CR questions; and (ii) a sampling error component. Unfortunately, it is not possible to identify these separate effects. We discuss how differences in student grade outcomes *across* assessments can be used to benchmark the size of MC-CR differences within the same assessment.

Empirical analysis of our large sample of students indicates that a switch to an all-MC format would result in grade changes that are in the “small” to moderate range when compared to grade changes that occur between assessments. This is noteworthy because the instructors in our classes made conscientious efforts to exploit the ability of CR questions to tap into higher levels of learning according to Bloom’s taxonomy (Bloom, 1956). On its face, this suggests that CR questions could be abandoned at relatively little cost in grading accuracy.

However, there are other arguments in favour of keeping CR questions. In particular, it has been suggested that students perceive a mix of CR and MC as “fairer” than an assessment composed exclusively of one or the other question type. Further, some instructors believe that CR questions encourage students to study harder.

Our study finds evidence to support both arguments. We surveyed a smaller number of students about their attitudes towards MC and CR questions. Our results confirm that students view mixed MC-CR assessments as being fairer than all-MC assessments. This holds true even for the subset of students who see themselves as personally advantaged by a certain question type. Further, approximately a third of students said they would study harder

if an assessment had CR questions, while a negligible number would study harder for MC questions, with the remainder saying they would be unaffected by question type. This suggests that the net effect would be greater study effort by students if assessments had CR questions.

For university instructors of introductory classes in economics, there is a frustrating lack of research that can guide them with respect to the choice of using CR questions versus the lower cost alternative of using all MC questions. This research attempts to fill that void. Our study is the first to estimate the effect that switching to all-MC assessments would have on students' grade outcomes. Our research suggests that moving to an all-MC format from a mixed question format would result in a relatively small change in students' grade outcomes. This could be interpreted as providing evidence in favour of adopting all-MC assessments. However, our survey of students' attitudes provides evidence that such a change would offend students' notions of fairness. Further, students' survey responses indicate that, on net, students are likely to study harder if assessments have some CR questions.

In conclusion, while our study sheds some light on the debate over MC versus CR questions, there is a need for additional research on this subject. An obvious direction for future research is to apply our methodology for measuring MC-CR grade differences to confirm whether similar results obtain in other classroom settings. With respect to students' notions of fairness, it would be interesting to know whether these might change if, for example, students could be convinced that grade outcomes would be little affected by a switch to all-MC assessments. Finally, the relationship between question format and students' study behaviour would be greatly illuminated by well-constructed field experiments. It is hoped that this study stimulates further research into this very practical subject.

REFERENCES

- Becker, W. E., & Johnston, C. (1999). The Relationship between Multiple Choice and Essay Response Questions in Assessing Economics Understanding. *Economic Record*, 75, 348-357.
- Bennett, R., E., Rock, D., A., & Wang, M. (1991). Equivalence of Free-Response and Multiple-Choice Items. *Journal of Educational Measurement*, 28(1), 77-92.
- Hickson, Stephen and Reed, W. Robert. "Do Constructed-Response and Multiple-Choice Questions Measure the Same Thing?" Working paper, University of Canterbury, May 2009.
- Kennedy, P. E., & Walstad, W. B. (1997). Combining Multiple-Choice and Constructed Response Test Scores: An Economists View. *Applied Measurement in Education*, 10(4), 359-375.
- Krieg, R., G., & Uyar, B. (2001). Student Performance in Business and Economic Statistics: Does Exam Structure Matter? *Journal of Economics and Finance*, 25(2), 229-241.
- Lukhele, R., Thissen, D., & Wainer, H. (1994). "On the Relative Value of Multiple-Choice, Constructed Response, and Examinee-Selected Items on Two Achievement Tests." *Journal of Educational Measurement*, 31(3), 234-250.
- Lumsden, K.G., & Scott, A (1987). The Economics Student Reexamined: Male-Female Differences in Comprehension. *Journal of Economic Education*, 18(4), 365-375.
- Thissen, D., Wainer, H., & Wang, X. (1994). Are Tests Comprising Both Multiple-Choice and Free-Response Items Necessarily Less Unidimensional Than Multiple-Choice Tests? An Analysis of Two Tests. *Journal of Educational Measurement*, 31, 113-123.
- Wainer, H. & Thissen, D. (1993). Combining multiple-choice and constructed response test scores: Towards a Marxist theory of test construction. *Applied Measurement in Education*, 6, 103-118.
- Walstad, W. B., & Becker, W. E. (1994). Achievement Differences on Multiple-Choice and Essay Tests in Economics. *American Economic Review*, 84, 193-196.

TABLE 1
Grade Distributions

<i>Grade (Points)^a</i>	<i>MC (Term Test) (1)</i>	<i>MC (Final Exam) (3)</i>	<i>CR (Term Test) (2)</i>	<i>CR (Final Exam) (4)</i>	<i>Historical Distribution of Final Grades (5)</i>
<i>A+ (9)</i>	11.1%	17.4%	4.0%	5.2%	3.4%
<i>A (8)</i>	15.5%	16.0%	4.2%	5.5%	4.7%
<i>A- (7)</i>	8.9%	8.6%	3.4%	5.6%	7.1%
<i>B+ (6)</i>	8.9%	16.3%	6.6%	8.8%	7.8%
<i>B (5)</i>	9.3%	7.8%	5.7%	7.6%	10.6%
<i>B- (4)</i>	17.2%	13.4%	9.5%	10.2%	10.9%
<i>C+ (3)</i>	6.4%	4.7%	6.9%	7.3%	12.6%
<i>C (2)</i>	6.3%	7.3%	10.1%	9.9%	19.7%
<i>C- (1)</i>	5.2%	2.9%	7.4%	6.8%	8.5%
<i>D (0)</i>	6.8%	3.0%	10.1%	8.3%	5.5%
<i>E (-1)</i>	4.3%	2.7%	32.2%	24.7%	9.3%
<i>Mean</i>	66.5	70.3	49.5	53.4	---
<i>Std. Dev.</i>	16.3	15.3	20.2	21.4	---

^a “Points” identifies the point allocation for each grade as used to calculate Grade Point Average (GPA).

NOTE: The bottom two rows in the table report the mean and standard deviation of raw assessment marks (adjusted to a 100-point scale) by assessment category. Columns (1) through (4) report the grade distributions that would arise from applying the UC grading schedule to the raw assessment marks. Column (5) reports the historical distribution of Final Grades given in all classes in our sample. A total of 15,508 observations are represented in the table.

TABLE 2
Distribution of MC Grades Conditional on CR Grades

		A+	A	A-	B+	B	B-	C+	C	C-	D	E	Sum
CR Grade (Term Test)	A+	28.5	17.2	18.0	12.4	7.1	6.7	6.7	3.4	0.0	0.0	0.0	100
	A	14.7	13.9	22.1	13.9	8.7	10.4	9.5	6.3	0.0	0.3	0.3	100
	A-	6.2	13.1	14.9	15.5	13.5	14.5	10.7	7.5	2.7	0.7	0.7	100
	B+	6.1	7.3	12.1	11.8	15.8	14.1	12.0	14.1	4.0	1.3	1.3	100
	B	2.6	5.0	9.5	9.5	15.1	15.0	15.8	17.4	5.0	2.3	2.8	100
	B-	2.9	4.5	6.9	10.0	11.4	11.3	16.0	20.5	8.3	3.3	4.9	100
	C+	0.8	3.5	5.5	8.1	12.4	11.2	14.3	25.1	8.2	4.1	6.8	100
	C	0.6	2.0	4.0	5.8	10.0	10.9	12.8	24.9	11.4	7.0	10.6	100
	C-	0.6	0.8	1.5	2.3	9.6	9.1	12.3	26.1	10.8	11.4	15.5	100
	D	0.0	0.7	0.7	2.1	5.2	7.3	9.9	23.7	17.1	14.8	18.5	100
	E	0.0	0.4	0.3	1.3	3.1	5.0	9.6	21.4	15.3	11.2	32.5	100
		A+	A	A-	B+	B	B-	C+	C	C-	D	E	Sum
CR Grade (Final Exam)	A+	35.6	20.2	14.6	14.6	7.1	4.1	2.6	0.0	0.0	0.0	1.1	100
	A	12.8	20.2	21.8	13.6	13.9	10.1	3.3	3.3	0.5	0.0	0.5	100
	A-	8.7	15.1	19.8	15.5	16.9	8.9	8.5	5.6	0.5	0.2	0.2	100
	B+	4.8	10.8	16.3	16.4	17.1	13.0	11.1	8.1	1.2	0.3	0.8	100
	B	1.7	4.5	9.1	13.2	19.5	16.1	15.7	16.3	2.6	0.7	0.6	100
	B-	2.1	2.6	7.4	9.9	13.3	15.6	19.7	20.3	5.3	1.9	1.9	100
	C+	1.0	2.2	3.9	6.3	13.4	15.2	16.5	24.3	9.8	3.8	3.6	100
	C	0.1	0.5	2.5	3.5	7.4	11.1	15.9	29.9	13.5	8.1	7.3	100
	C-	0.5	0.2	1.1	2.0	3.0	6.5	11.5	30.5	14.9	12.9	16.9	100
	D	0.2	0.2	0.5	1.4	3.1	5.6	8.5	27.7	18.1	14.1	20.7	100
	E	0.0	0.1	0.3	0.6	0.8	2.6	4.3	16.1	14.3	13.2	47.6	100

NOTE: Numbers in tables are percentages and should be interpreted as follows: 28.5 percent of all Term Test takers who received an A+ on their CR component also earned an A+ on the MC component of the test (after norming the respective grade distributions to the historical averages in TABLE 1). 17.2 percent of all Term Test takers who received an A+ on their CR component earned an A on the MC component of that assessment. And so on. As the probabilities are conditional on grade received on the CR component, all row probabilities sum to 100%.

TABLE 3
Comparison of Grade Differences for Different Combinations of MC and CR Assessments

	Negative Difference of One Letter Grade or More	Difference of Less Than One Letter Grade	Positive Difference of One Letter Grade or More
<i>MC-CR (Term)</i>	16.2%	67.0%	16.7%
<i>MC-CR (Final)</i>	11.4%	77.3%	11.3%
<i>MC(Term) - MC(Final)</i>	16.8%	66.6%	16.5%
<i>CR(Term) - CR(Final)</i>	12.6%	74.9%	12.5%

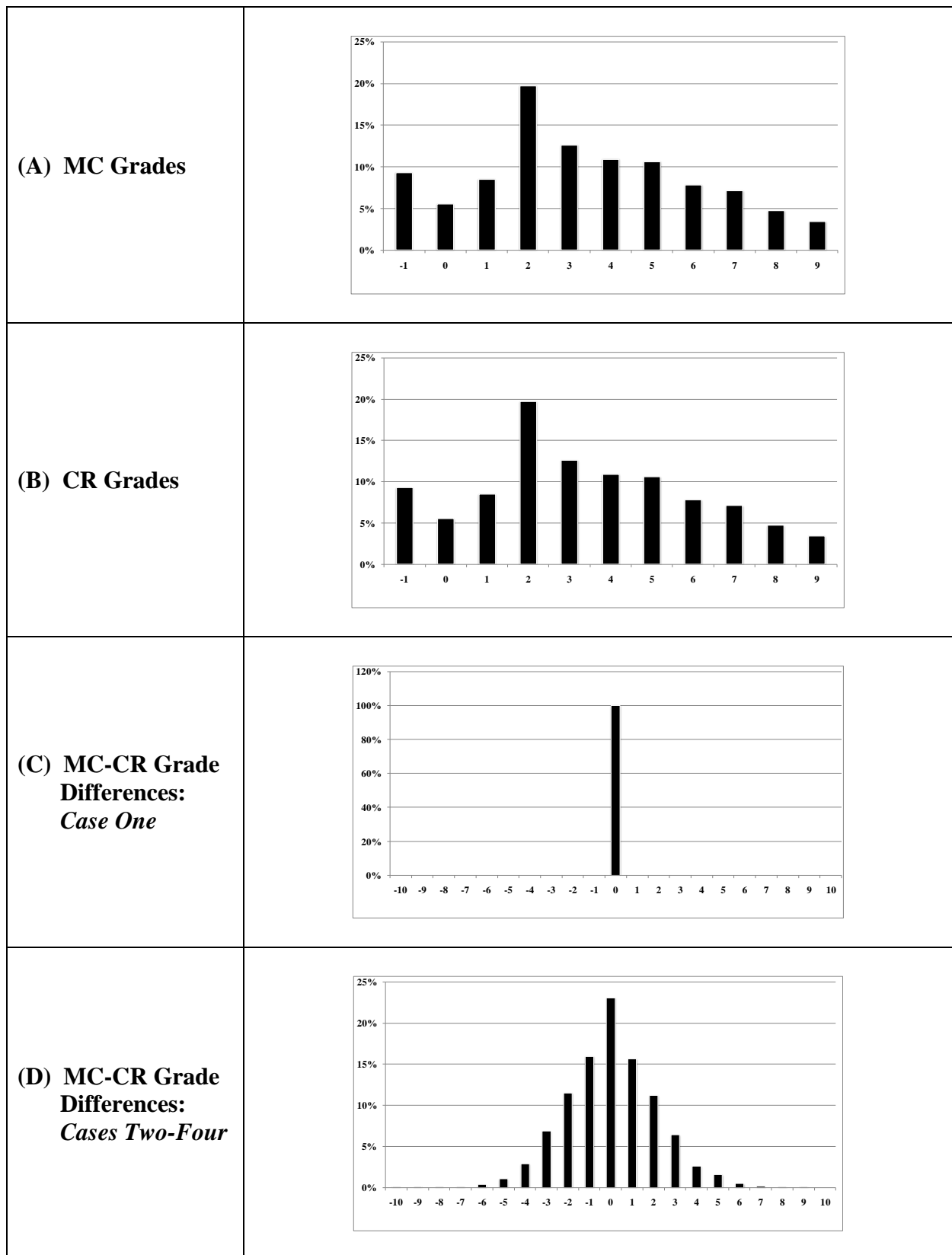
NOTE: The first two rows use the data from TABLE 2 to calculate differences between CR and MC grades on the same assessment. Examples of a “Negative Difference of One-Letter Grade” are CR Grade = A- and MC Grade = B-; CR Grade = B+ and MC Grade = C+; etc. A “Positive Difference of One-Letter Grade” is defined similarly. The last two rows report grade differences between (i) MC grades on the Term Test and Final Exam and (ii) CR grades on the Term Test and Final Exam, respectively.

TABLE 4
Survey of Student Attitudes About MC and CR Questions

QUESTION	PERCENT
Q1: For a test or exam, which format do you think would give you the highest rank in class? <i>NOTE: This question is not about which format gives you the highest score. Raw scores for multiple choice are usually higher due to guessing. This is about your rank in class.</i>	
A test where all the questions are multiple choice.	14.5
A test where all the questions require a written answer.	15.3
A test with a mixture of multiple choice and written answer questions.	59.5
The test format would not make any difference to my rank in class.	10.7
TOTAL	100.0
Q2: For a test or exam, which format do you think is fairest for students?	
A test where all the questions are multiple choice.	3.1
A test where all the questions require a written answer.	6.9
A test where there are both multiple choice and written answer questions.	79.2
All of the above are equally fair.	10.8
TOTAL	100.0
Q2.A For a test or exam, which format do you think is fairest for students? (Restricted to the 39 students who expressed a preference for all-MC or all-CR questions in Q1)	
A test where all the questions are multiple choice.	5.4
A test where all the questions require a written answer.	16.2
A test where there are both multiple choice and written answer questions.	73.0
All of the above are equally fair.	10.8
TOTAL	100.0
Q3: Which one of these best describes you?	
I would study harder for a test if all the questions were multiple choice.	5.3
I would study harder for a test if all the questions required a written answer.	35.1
The test format makes no difference to how hard I would study.	59.5
TOTAL	100.0

TOTAL NUMBER OF RESPONSES = 131

FIGURE 1
Distributions of MC Grades, CR Grades, and MC-CR Grade Differences



NOTE: Panels (A) and (B) report the distributions of MC and CR grades for a given assessment, where grades are measured in points as reported in TABLE 1. Each assessment is “normed” to have the same distribution as the historical distribution of Final Grades for all student observations in our sample (cf. Column 5 in TABLE 1). Panels (C) and (D) report the distributions of MC-CR grade point differences that would arise under different cases. For example, a student who earned an A- (7 points) on the MC component of an assessment, and a B- (4 points) on the CR component of that assessment would have a grade difference of -3 points. The four cases characterize: (i) no sampling error and no “systematic differences;” (ii) no sampling error and “systematic differences;” (iii) sampling error and no “systematic differences;” and (iv) both sampling error and “systematic differences.” The four cases are discussed in greater detail in the text.